Optimal Combination of Glycan-Based Serum Diagnostic Markers Which Maximize AUC

Marko I. Vuskovic and Haofei Fang
Department of Computer Science, San Diego State University
San Diego, 92182 CA, USA

Harvey I. Pass

and

Margaret E. Huflejt
Department of Cardiothoracic Surgery, New York University
New York, 10016 NY, USA

ABSTRACT

Recently a new high-throughput biomarker discovery platform based on printed glycan arrays (PGA) has emerged. PGAs are similar to DNA arrays but contain deposits of various carbohydrate structures (glycans) instead of spotted DNAs. PGA-based biomarker discovery for the early detection, diagnosis and prognosis of human malignancies is based on the response of the immune system as measured by the level of binding of anti-glycan antibodies from human serum to the glycans on the array. Since the PGA offer a multitude of markers which can have moderate individual diagnostic power they can be combined in order to achieve maximal classification precision assessed by the popular performance measure area under the ROC curve (AUC). This paper presents an empirical analysis of several combination approaches including those that are specifically designed to maximize the AUC and those that are not, such as Fisher Linear Discriminant, Support Vector Machines and Generalized Linear Model. The analysis is performed on real-life PGA data from three pilot studies involving malignant mesothelioma, lung cancer and ovarian cancer.

Keywords: Printed glycan arrays, biomarkers, AUC, optimal combination of biomarkers, mesothelioma

1. INTRODUCTION

In the last five years a new biomarker-discovery platform has emerged based on glycan arrays [4], that has some advantages over nucleic acid-based and other platforms. The printed glycan arrays (PGA) are similar to DNA microarrays, but contain deposits of various carbohydrate structures (glycans) instead of spotted DNAs. Most of these glycans can be found on the surfaces of normal human cells, human cancer cells, and on the surfaces of many human infectious agents such as bacteria, viruses, and other pathogenic microorganisms. Transformation of cells from healthy to pre-malignant and malignant is associated with the appearance of abnormal glycosylation on proteins and lipids presented on the surface of these cells. The malignancy-related abnormal glycans are called tumor-associated carbohydrate antigens (TACA), [10]. There is growing evidence [1] that numerous TACAs are immunogenic, and that the human immune system can generate antibodies against them. Since multiple glycans arrayed on PGAs are either known TACAs or closely related structures, the antibodies present in human sera that bind to glycans on PGA can indicate the status of response of the immune system to human malignancies [14,15,16]. A prototype of PGA with a library of 200 glycan structures was built at Scripps Research Institute, La Jolla, California, under the auspices of the Consortium of Functional Glycomics (CFG), [3]. Further development and standardization of the PGA with 211 glycans was conducted at Cellexicon, Inc., La Jolla, in collaboration with Nicolai Bovin of Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, of the Russian Academy of Sciences, Moscow, Russia. The "second generation" of PGA was used in several pilot studies on early detection of cancer and cancer risk supported by the National Cancer Institute. Research and improvement of PGA technology and its relevance in diagnostic and prognostic applications is currently continuing in the Tumor Glycome Research Group at the Thoracic Surgical Laboratory at the New York University, School of Medicine, in collaboration with the Shemyakin-Ovchinnikov Institute.

The advantages of potential PGA-based serum test for early detection of cancer and cancer risk can be summarized as follows: (a) minimal invasiveness of serum sampling; (b) minimal sampling variability, in contrast to well known heterogeneity of solid tissue samples; (c) stability of antibodies, (d) low cost associated with technology; (e) low labor intensity and short duration of the test; (f) broad scope of the test, i.e. the test doesn't have to be narrowly targeted to a particular disease, e.g. cancer type. All these advantages make the PGA platform attractive for early detection of disease and for the potential application in screening of general population.

This study is motivated by the need for combination of several biomarkers due to relatively moderate discriminatory power of individual glycans caused by still limited glycan library of the early generation of PGA arrays, and due to relatively small sample sizes. During the extensive experimentation in pilot studies we have recognized the importance of the area under the ROC curve (AUC) as a consistent and robust performance measure of the discriminatory power of classifiers. Therefore we have explored some recently proposed combination approaches that maximize the AUC value, as well as the relevance of popular classifiers such as Fisher Linear Discriminant (FLD), Support Vector Machine (SVM) and Generalized Linear Model (GLM) in the context of maximization of AUC. In addition to these approaches we have also added Ant Colony Optimization (ACO) [6], which is recently gaining popularity in biomarker discovery [22, 25]. Our version of ACO uses AUC as fitness function.

It is important to mention, that the goal of this paper is not to propose PGA-based putative signatures, but rather to evaluate various combination methods in the light of AUC and PGA-based predictors.

2. PRINTED GLYCAN ARRAYS

A printed glycan array (PGA) consists of a glass slide coated with a chemically reactive surface on which various glycans are covalently attached using standard amino-coupling chemistry and contact printing technology [3]. A PGA slide contains several sub-arrays of the entire, currently available glycan library in form of microscopic glycan deposits of size about 80 microns. The version of the PGA used to generate data presented in this paper has two concentrations of glycans (10 and 50 μM) and eight replicates for each concentration, thus resulting in an array of 16 sub-arrays, each containing 211 deposits of unique glycan structures, and biotin spots used as a printing control.

The measurement of binding of human anti-glycan antibodies (AGA) to arrayed glycans is achieved as described in [17]. The PGA slide is first incubated with the subject's serum, allowing the binding of serum antibodies to glycans in PGA deposits. Serum IgG, IgM and IgA immunoglobulins bound to printed glycans are visualized simultaneously with the "combo" biotinylated secondary goat anti human IgG, IgM and IgA antibodies (Pierce Biotechnology, Inc., Rockford, IL) and streptavidin-Alexa^{555} (Invitrogen/Molecular Probes, Carlsbad, CA). Fluorescence signal intensities that correspond to antibodies bound to printed glycans are scanned at 90% laser power, and quantified with ImaGene software (BioDiscovery, Inc., El Segundo, CA). The total relative fluorescence signal intensity values (appx. range: 1,000 – 32,000,000 Relative Fluorescence Units) are used for further data preprocessing and analyses. The preprocessing included signal screening for noise, normalization and normality transformation.

The population size of the initial studies involving the early generation of PGAs with 211 glycans on the array and which are used in this study is shown in **Table** 1. The choice of these studies is made for their diversity in terms of resubstitution AUC and imbalance of control and case samples.

Table 1: Serum samples used in pilot PGA-based studies

Study	*Source*	*Control sample* (n_1)	*Case sample* (n_2)
Mesothelioma	NYUSM	Asbestos exposed (65)	Malignant mesothelioma (50)
Lung cancer	NYUSM	Smokers (49)	Adenocarcinoma (46)
Ovarian cancer	MDACC	Healthy donors (106)	Early aggressive o.c., stage I/II (21)

NYUSM – Department of Cardiothoracic Surgery, School of Medicine, NYU (Dr. Harvey I. Pass)
MDACC – MD Anderson Cancer Center, Univ. of Texas (Dr. Karen H. Lu)

3. OPTIMAL LINEAR COMBINATION WHICH MAXIMIZES AUC

In this section we briefly discuss advantages of using AUC as the performance measure in evaluation of combined multiple marker tests, and formulate the optimization problem.

Suppose a predictor matrix $\boldsymbol{X} = [x_{ij}]$, $i = 1,2,\ldots,n$, $j = 1,2,\ldots,d$ where x_{ij} is the marker value for i-th patient and j-th marker. In our case x_{ij} represents a continuous measurement of the intensity of binding of human antibodies of patient i against glycan structure j deposited on the PGA array. These values are usually normalized and transformed before the diagnostic analysis. The matrix $\boldsymbol{X}$ is associated with a column vector $\boldsymbol{y} = [y_i]$ where y_i are labels for control ($y_i = 1$) and case observations ($y_i = 2$). In many practical situations the individual markers do not provide sufficient discriminatory power, and they have to be combined. Simple approach to this is linear combination $\boldsymbol{z} = \boldsymbol{Xw}$, where $\boldsymbol{z}$ is column vector of combined markers and $\boldsymbol{w}$ is column vector of combination coefficients, usually in normalized form, i.e. $||\boldsymbol{w}|| = 1$. The diagnostic test of an unknown patient, whose marker values are $\boldsymbol{u} = [u_1, u_2, \ldots, u_d]$ can be achieved by testing the sign of $\boldsymbol{u}\,\boldsymbol{w} + w_o$, where w_o is a decision point determined, as well as $\boldsymbol{w}$, from the training set ($\boldsymbol{X}$, $\boldsymbol{y}$). There are several popular approaches to determine the vector $\boldsymbol{w}$, for example Fisher Linear Discriminant (FLD), Generalized Linear Model (GLM), Support Vector Machines (SVM), which are based on essentially similar optimization criteria, such as Mahalanobis distance between sample means, likelihood of odds to belonging to one of the samples, and margin between samples, respectively. The performance of the classifier ($\boldsymbol{w}$, w_o) is often measured by the accuracy computed for a test set:

$$Acc = f_1(\boldsymbol{Tw}, \boldsymbol{y}_t, w_o) = \frac{TN + TP}{n_1 + n_2}, \qquad (1)$$

where $\boldsymbol{T}$ is matrix of markers of the test set (can be replaced with $\boldsymbol{X}$ for resubstitution accuracy), $\boldsymbol{y_t}$ are labels for the test set, n_1 and n_2 are the numbers of control and case observations in $\boldsymbol{T}$, while TN (true negatives) and TP (true positives) are numbers of correctly classified control and case observations, using the discriminant function $\boldsymbol{t}\,\boldsymbol{w} + w_o$, where $\boldsymbol{t}$ is a row of $\boldsymbol{T}$.

Although this performance measure is somewhat straightforward and natural, it indeed has several drawbacks: (a) the measure depends on the decision point w_o, (b) the measure is largely affected by the sample imbalance, and (c) the measure has low resolution, specially in case of smaller samples, which notoriously occurs in cross-validation tests. These reasons are discussed in [7] and [18].

An alternative measure that doesn't have these drawbacks is the Area Under the ROC curve (AUC). The ROC curve (Receiver Operating Characteristic curve) is introduced [24] as a graphical tool to evaluate discriminatory accuracy of binary classifiers for various decision points. A single number measure for a family of classifiers which entirely eliminates w_o, the area under the ROC curve, was proposed by [12,5], and is used since as a standard in biomedicine and bioinformatics. The best possible value of AUC (complete discrimination) yields AUC = 1, while AUC = 0.5 means that the classifier has no discriminative power). In addition to the advantages mentioned above, the AUC also captures the ranking ability of the classifier, which is an important notion even more fundamental than classification [8].

As proposed by [11], the AUC value can be computed for a given training, or test set, ($\boldsymbol{X}$, $\boldsymbol{y}$) and a given combination vector $\boldsymbol{w}$, without previously deriving the ROC curve:

$$AUC = f_2(\boldsymbol{Tw}, \mathbf{y}_t) = \frac{S_2 - n_2(n_2+1)/2}{n_1 n_2} \ , \qquad (2)$$

where S_2 is sum of ranks of rows of $\boldsymbol{Tw}$ taken for case observations [11].

The optimal combination vector can be now estimated from the training set:

$$\hat{\boldsymbol{w}} = \arg\max_{\boldsymbol{w}} f_2(\boldsymbol{Xw}, \boldsymbol{y}) \qquad (3)$$

Since (2) has no closed form solution, the optimization must be performed by some global optimization approach, such as Genetic Algorithm. Another approach would be to develop a closed form of (2) under some simplifying assumptions about the distributions of control and case samples. This was shown in [23], under the assumption that the two samples are normally distributed with arbitrary covariance matrices. After using the maximum likelihood estimators for the covariance matrices and means, the estimate of the optimal combination vector can be expressed in terms of sample covariances and means:

$$\hat{\boldsymbol{w}} = (\mathrm{cov}(\boldsymbol{X}_1) + \mathrm{cov}(\boldsymbol{X}_2))^{-1}(mean(\boldsymbol{X}_2) - mean(\boldsymbol{X}_1))^T \quad (4)$$

where $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are predictor matrices for control and case samples respectively. We will refer to this method of optimization as Optimal Combination under Normality assumption (OCN).

Computation of optimal combination $\boldsymbol{w}$ which maximizes AUC under less restrictive assumption than normality is investigated by [21,19,20,13]. The approach is based on Generalized Linear Model assumption. The AUC can be generally expressed [2]:

$$AUC(\boldsymbol{w}) = \Pr(\boldsymbol{x}_j\boldsymbol{w} > \boldsymbol{x}_i\boldsymbol{w}), \ i \in I_1, j \in I_2 \qquad (5)$$

where $\boldsymbol{x}_k$ is row vector of markers for patient k from the training, or test samples, I_1 and I_2 are sets of row indices of matrix $\boldsymbol{X}$ which correspond to control and case samples respectively. Equation (5) can be extended to:

$$AUC(\boldsymbol{w}) = \frac{1}{n_1 n_2} \sum_{i \in I_1} \sum_{j \in I_2} I(\boldsymbol{x}_j\boldsymbol{w} - \boldsymbol{x}_i\boldsymbol{w} > 0) \qquad (6)$$

where I is indicator function. The function (6) is clearly discrete and can not be used by some convenient numerical optimization method. Ma and Huang [19,20] have therefore suggested replacing the discrete indicator function with some smooth, monotonically increasing function, such as sigmoid function $s(z) = 1/(1+\exp(-z/\sigma))$. The optimization of $AUC(\mathbf{w})$ can then be formulated as:

$$\hat{\boldsymbol{w}} = \arg\max_{\boldsymbol{w}} \sum_{i \in I_1} \sum_{j \in I_2} s(\boldsymbol{x}_j\boldsymbol{w} - \boldsymbol{x}_i\boldsymbol{w}) \qquad (7)$$

To solve (7) we have used the Newton-Raphson method which converges very rapidly for a proper choice of parameter σ, which was kept constant throughout the optimization process. This algorithm will be referred to as Optimal Combination under GLM assumption (OCG). The implementation of the algorithm is presented in Appendix. As shown in the following section, the execution speed of this algorithm is close to FLD algorithm.

4. EMPIRICAL EVALUATION OF APPROACHES

The comparison of various approaches for finding the optimal combination vector $\boldsymbol{w}$ will be done with the reference to the "best possible approach" obtained with a global optimization. For this purpose we have used genetic algorithm (GA) introduced by [9]. We start with the mesothelioma dataset, for which the library of glycans on PGA appeared to be most complete. For feature selection we have used the univariate non-parametric Wilcoxon-Mann-Whitney rank sum test. The multivariate approach for feature selection is avoided in this study due to relatively small sample sizes and the risk of over-fitting. The five most discriminatory glycans are indicated in **Table** 2. The table also shows the individual AUC values obtained by equation (2). Clearly, the individual AUC values are relatively low which justifies the need for combination of multiple markers.

Table2: Glycan structures for top five glycans determined by Wilcoxon-Mann-Whitney rank sum test, for Mesothelioma study

Glycan Structure	*AUC*
Neu5Acα2-3Galβ1-4Glcβ-sp	0.7274
(Neu5Acα2-8)3-sp	0.6923
GlcNAcβ1-6GalNAcα-sp	0.6889
GlcNAcβ1-4(GlcNAcβ1-6) GalNAcα-sp	0.6868
Galβ1-3GlcNAcβ1-3Galβ1-4Glcβ-sp	0.6548

The observed (training) AUC values for combination of markers for various approaches, AUC_O, are shown in **Table** 3. As seen, the best training performance besides the global optimization is obtained for OCG and GLM. These approaches are also very efficient in terms of execution time (ET). Although the differences are relatively small, it is desirable to establish their statistical significance. For this purpose we have used paired bootstrap with replacement with 1000 replications. The box plots of the corresponding empirical distributions of differences are shown in **Figure** 1, and the empirical medians are summarized in **Table** 3 (for the variations of medians are used median absolute deviations, MAD). The figure suggests that the differences are statistically significant. This was also verified with paired ANOVA test for all six groups, and with non-parametric Wilcoxon signed-rank test performed for all combinations of pairs of approaches. In both cases all p-values were close to zero.

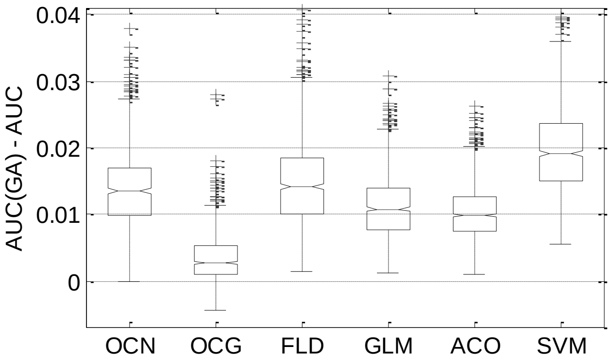

Figure 1: Empirical distributions of paired differences between AUC obtained with GA and other approaches generated by bootstrap with replacement using 1000 replications in Mesothelioma study.

Figure 1 and **Table** 3 are concerned with the resubstitution AUC values. In order to evaluate the generalization power of the approaches considered above we applied the unbiased repeated 10-fold cross-validation with 100 repetitions. The results are shown in **Figure** 2. The cross-validation test was applied for various set sizes which range from a single glycan up to combination of seven glycans. The cross-validated AUC value for all approaches has reached the maximum at 5 to 6 glycans. The curves begin to drop after the glycan sets become larger than six (not shown here) which is a consequence of over-fitting. The cross-validation results are summarized in **Table** 3. It is interesting to notice that SVM has demonstrated the best performance, which is however only slightly above GLM (for $d = 5$) .

The box-plots for empirical distributions of paired differences between GA and other approaches for resubstitution AUC for the other two studies listed in **Table** 1 are shown in **Figures** 3 and 4. The figures again indicate small differences, which are still statistically significant judging by ANOVA and Wilcoxon signed-rank tests. The ranking in resubstitution performance is basically similar as for the Mesothelioma study.

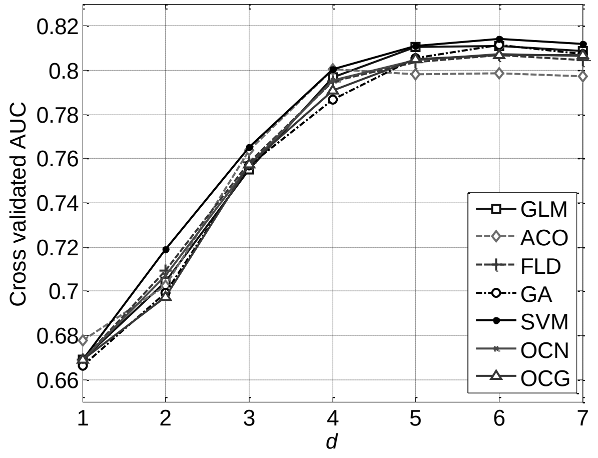

Figure 2: Cross-validated AUC values for various number of markers and various approaches in Mesothelioma study

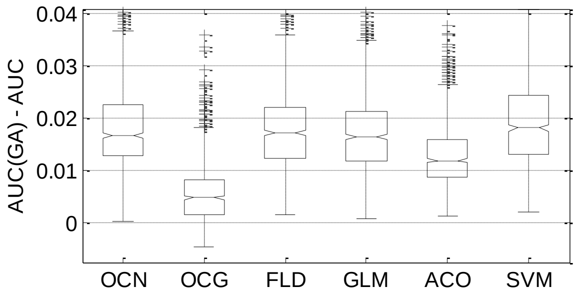

Figure 3: Empirical distributions of paired differences between AUC for lung cancer

Table 3: Observed, bootstrapped and cross-validated AUC for Mesothelioma study ($n_1 = 65$, $n_2 = 50$, $d = 5$)

Method	*AUCo*	*Median AUC*	*AUC$_{CV}$*	*ET (sec)*
OCN	0.8582	0.8729 ± 0.019	0.8046	0.036
OCG	0.8649	0.8837 ± 0.018	0.8051	0.114
FLD	0.8566	0.8717 ± 0.019	0.8039	0.105
GLM	0.8637	0.8754 ± 0.019	0.8106	0.123
ACO	0.8559	0.8769 ± 0.019	0.7980	2.744
SVM	0.8578	0.8677 ± 0.018	0.8112	0.324
GA	0.8665	0.8877 ± 0.018	0.8054	3.261

OCN – optimal combination under normality assumption (equation (4))
OCG – optimal combination under GLM assumption (equation (7))
FLD – Fisher Linear Discriminant
GLM – Generalized Linear Model (without interaction terms)
ACO – Ant Colony Optimization
SVM – Support Vector Machine
GA – Genetic Algorithm (global optimization)

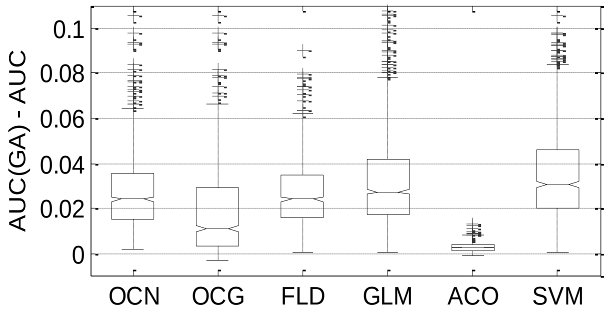

Figure 4: Empirical distributions of paired differences between AUC for ovarian cancer

5. CONCLUSION

The goal of this study was to empirically evaluate several approaches for combining the multiple test markers based on printed glycan array data obtained from three pilot studies, and to conclude whether any of these methods is particularly superior in light of the maximization of AUC value. Three of these methods (GA, OCN and OCG) are specifically designed to optimize AUC, while the others (FLD, ACO, GLM and SVM) are not designed for this purpose but are very popular in diagnostic classifiers. The paired bootstrap tests have shown that there is a small difference in performance of resubstituted AUC value, in favor of GA and OCG, but the difference is not substantial enough to disfavor approaches as GLM and SVM. Moreover, the cross-validated AUC performance evaluated on mesothelioma study has shown that the SVM and GLM provide slightly better predictive precision and predictive AUC value than the approaches specifically designed to optimize AUC. The relevance of these findings will be reexamined in the near future with the next generation of PGA arrays with 300 and 400 glycans, and on larger serum populations.

ACKNOWLEDGEMENT

The work presented in this paper was partially supported by NCI grant ‘Discovery and Clinical Validation of Cancer Biomarkers Using Printed Glycan Array’ (Grant No.: U01CA128526 to M.E.H) and NCI/EDRN grant ‘American/Australian Mesothelioma Consortium’ (Grant No.: U01CA111295 to H.I.P.). The PGAs were printed and developed at Cellexicon, Inc., La Jolla, CA. Most of the glycans used in PGA were synthesized and purified in the Carbohydrate Chemistry Laboratory at Shemyakin Institute of Bioorganic Chemistry, Moscow, Russia, under the support of the grant from RAS Presidium Program "Molecular and Cell Biology" (under supervision of Prof. Nicolai Bovin).

REFERENCES

[1] C.A. Aarnoudse, J.J. Garcia Vallejo, E. Saeland, Y. van Kooyk, “Recognition of tumor glycans by antigen presenting cells”, **Curr. Opin. Immunol**. 18:105-111, 2006.

[2] D. Bamber, “The Area Above the Ordinal Dominance Graph and the Area Below the Receiver Operating Characteristic Graph”, **Journal of Mathematical Psychology**, 12, 1975, pp. 387-415.

[3] O. Blixt, Head, S.T. Mondala, C. Scanlan, M.E. Huflejt, R. Alvarez, M.C. Bryan, F. Fazio, D. Caralese, J. Stevens, N. Razi, D.J. Stevens, J.J. Skehel, I. van Die, D.R. Burton, I.A. Wilson, R. Cummings, N. Bovin, C-H. Wong, and J.C. Paulson, “Printed covalent glycan array for ligand profiling of diverse glycan binding proteins”, **PNAS**, 101, 2004, pp.17033-17038

[4] N.V. Bovin, and M.E. Huflejt, “Unlimited Glycochip”, **Trends in Glycoscience and Glycotechnology**, 20: 245–258, 2008

[5] A.P. Bradley, “The Use of the Area under the ROC Curve in the Evaluation of Machine Learning Algorithm”. **Pattern Recognition**, Vol. 30, No. 7, 1997, pp. 1145-1159.

[6] M. Dorigo and L.M. Gambardella “Ant Colony System : A Cooperative Learning Approach to the Traveling Salesman Problem”, **IEEE Transactions Evolutionary Computation**, Vol. 1, 1997, pp 53-66.

[7] T. Fawcett, “ROC Graphs: Notes and Practical Considerations for Researchers”, **Technical Report**, HPL-2003-4, Intelligent Enterprise Technologies Laboratory, HP Laboratories Palo Alto. , 2003

[8] P. Flach, “Tutorial on The Many Faces of ROC Analysis in Machine Learning”, **21st International Conference on Machine Learning (ICML)**, 2004.

[9] D.E. Goldberg, “Genetic Algorithms in Search, Optimization and Machine Learning”, **New York Addison-Wesley**, 1989.

[10] S. Hakomori, “Glycosylation defining cancer malignancy: new wine in an old bottle” **PNAS** 99:10231, 2002.

[11] D. Hand and R. Till “A simple generalization of the area under the ROC curve for multiple class classification problem”, **Machine Learning**, Vol. 45, 2001, pp 171-186.

[12] Hanley, and B. J. McNeil, “The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve”, **Radiology,Vol**. 143, 1982, pp. 29-36.

[13] X. Huang, G. Qin and Y. Fang, “Optimal Combination of Diagnostic Tests Based on AUC”, ***Biometrics (Epub)***, June 16 2010,

[14] M.E. Huflejt, M. Cristofanilli, L.E. Shaw, J.M. Reuben, H.A. Fritsche, G.N. Hortobagyi and O. Blixt, “Detection of neoplasia-specific clusters of anti-glycan antibodies in sera of breast cancer patients using a novel glycan array”, **Proc. Amer. Assoc. Cancer Res**, 2005a, (46) .

[15] M.E. Huflejt, M. Vuskovic, O. Blixt, H. Xu, L.E. Shaw, J.M. Reuben, H.M. Kuerer and M. Cristofanilli, “Glycan array identifies specific signatures of anti-glycan auto antibodies in sera of breast cancer Patients: diagnostic, prognostic and therapeutic opportunities”. **28th Annual San Antonio Breast Cancer Symposium**, San Antonio, TX. *Breast Cancer Res. Treat.* 94: S85, 2005b.

[16] M.E. Hufleit, O. Blixt, M. Vuskovic, H. Xu, L.E. Shaw, J.M. Reuben, H.M. Kuerer and M. Christofanilli, “Anti-glycan autoantibodies as markers of malignancy status”, **5th Internat. Symposium on Minimal Residual Cancer**, San Francisco, September 11-14, 2005c,.

[17] M. E. Huflejt, M. Vuskovic, D. Vasiliu, H. Xu, P. Obukhova, N. Shilova, A. Tuzikov, O. Galanina, B. Arun, K. Lu, N. Bovin, “Anti-carbohydrate antibodies of normal sera: Findings, surprises, and challenges”, **Molecular Immunology**, 46, 2009, pp. 3037-3049.

[18] C.X. Ling, J. Huang, H. Zhang, “AUC: a statistically consistent and more discriminating measure than accuracy”, **Proceedings of the Eighteenth International Joint Conference of Artificial Intelligence (IJCAI)**, 2003.

[19] S. Ma, and J. Huang, “Regularized ROC method for disease classification and biomarker selection with microarray data”, **Bioinformatics**. Vol. 21 no. 24, 2005, pp 4356–4362.

[20] S. Ma and J. Huang, "Combining Multiple Markers for Classification Using ROC", **Biometrics** Vol. 63, 2007, pp. 751–757.

[21] M.S. Pepe and M.L. Thompson, "Combining Diagnostic Test Results to Increase Accuracy", **Biostatistics** Vo. 1, No. 2, 2000, pp. 123-140.

[22] H.W. Ressom, R.S. Varghese, S.K. Drake, G.L. Hortin, M. Abdel-Hamid, C.A. Loffredo, and R. Goldman, "Peak selection from MALDI-TOF mass spectra using ant colony optimization", **Bioinformatics** 23(5):619-626, 2007.

[23] J.Q. Su, and J.S. Liu , "Linear Combinations of Multiple Diagnostic Markers", **Journal of the American Statistical Association**, Vol. 88, No. 424 (Theory and Methods), 1993.

[24] J. A. Swets and R.M. Pickett, "Evaluation of Diagnostic Systems: Methods from Signal Detection Theory", **Academic Press**, New York, 1992.

[25] Y. Wang, X. Liu, K. Robbins, and R. Rekaya, "Ant- Epi-Seeker: Detecting Epistatic Interactions for Case-Control Studies Using a Two-Stage Ant Colony Optimization Algorithm", **BMC Research Notes**, 3:117, 2010.

APPENDIX

The OCG algorithms based on equation (7) is implemented in MATLAB which supports vectorized functions and rapid array index manipulation. In order to make the pseudocode below more concise we will define some basic overloaded functions and index manipulation operators, using simple examples.

Suppose vectors $\boldsymbol{a} = [a_1, a_2, a_3]$ and $\boldsymbol{b} = [b_1, b_2, b_3]$,

then: $\exp(\boldsymbol{a}) = [\exp(a_1), \exp(a_2), \exp(a_3)]$,

$$\boldsymbol{a} + \alpha = [a_1 + \alpha, a_1 + \alpha, a_1 + \alpha],$$

$$\boldsymbol{a} \bullet\!* \boldsymbol{b} = [a_1 b_1, a_2 b_2, a_3 a_3],$$

$$\boldsymbol{a} \bullet\!/ \boldsymbol{b} = [a_1 / b_1, a_2 / b_2, a_3 / b_3],$$

$$sum(\boldsymbol{a}) = a_1 + a_2 + a_3,\ rep(\boldsymbol{a}, 4) = [\boldsymbol{a} \mid \boldsymbol{a} \mid \boldsymbol{a} \mid \boldsymbol{a}]$$

$$rep(\boldsymbol{a}^T, 4) = [\boldsymbol{a}^T \mid \boldsymbol{a}^T \mid \boldsymbol{a}^T \mid \boldsymbol{a}^T]$$

$$row(\boldsymbol{A}) = [row_1(\boldsymbol{A}) \mid row_2(\boldsymbol{A}) \ldots \mid row_3(\boldsymbol{A})],$$

$$\boldsymbol{a}([2,2,1,3]) = [a_2, a_2, a_1, a_3].$$

Also $A(i,*)$ and $A(*, j)$ denote *i*-th row and *j*-th column of matrix ***A*** respectively**.** Finally, we need notion of constant vectors $\boldsymbol{e}(n) = [1,1,...,1]$ (*n* repeated 1's), and $\boldsymbol{c}(n) = [1, 2, ..., n]$. Using these definitions the gradient ***g*** and Hessian matrix ***H*** of *AUC(**w**)* , as well as the optimization process are implemented as follows:

$$\boldsymbol{k}_1 = row(rep(\boldsymbol{c}(n_1), n_2);$$

$$\boldsymbol{k}_2 = rep(\boldsymbol{c}(n_2)^T, n_1);$$

{***k***1 and ***k***2 are (*n1n2*)-element constant index vectors which are generated before the optimization}

$$\boldsymbol{z}_1 = \boldsymbol{X}_1 \boldsymbol{w};\ \ \boldsymbol{z}_2 = \boldsymbol{X}_2 \boldsymbol{w};$$

{***X***1 and ***X***2 are sub matrices of ***X*** which correspond to control and case samples}

$$\boldsymbol{z} = \boldsymbol{z}_2(\boldsymbol{k}_2) - \boldsymbol{z}_1(\boldsymbol{k}_1);$$

$$\boldsymbol{p} = \boldsymbol{e}(n_1 n_2) \bullet\!/ (1 + \exp(\boldsymbol{z}));$$

{The vector of sigmoid values}

$$J = sum(\boldsymbol{p});$$

{The objective function *n1n2 AUC*(***w***)}

$$\boldsymbol{Z} = \boldsymbol{X}_2(\boldsymbol{k}_2, *) - \boldsymbol{X}_1(\boldsymbol{k}_1, *);$$

$\boldsymbol{q} = \boldsymbol{p} \bullet\!* (1 - \boldsymbol{p}) / \sigma;$ {First derivative of ***p*}**

$$\boldsymbol{Q} = rep(\boldsymbol{q}, d);$$

$\boldsymbol{g} = sum(\boldsymbol{Q} \bullet\!* \boldsymbol{Z});$ {The gradient}

$\boldsymbol{r} = \boldsymbol{q} \bullet\!* (1 - 2\boldsymbol{p}) / \sigma;$ {The second derivative of ***p*}**

$$H_{ij} = sum(\boldsymbol{r} \bullet\!* \boldsymbol{Z}(*, i) \bullet\!* \boldsymbol{Z}(*, j));$$

{Elements of the Hessian matrix}

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - \boldsymbol{H}(\boldsymbol{w}_k)^{-1} \boldsymbol{g}(\boldsymbol{w}_k);$$

{Iterative process, the initial value ***w***1 is determined by OCN , eq. (4)}

$| J(\boldsymbol{w}_{k+1}) - J(\boldsymbol{w}_{k+1}) | \le \varepsilon$ {The stopping criteria}

The equations above could have been implemented in standard vector notation, which would however result in a couple of hundred times slower execution.